Statistical Segmentation of Biological Sequences

CHEONG Siew Ann

Division of Physics and Applied Physics,
School of Physical and Mathematical Sciences,
Nanyang Technological University

2008 BIRC Workshop on Advances in Bioinformatics
16 February 2008

Acknowledgments

- Postdoctoral work in collaboration with:

Christopher R. Myers
Center for Advanced Computing,
Cornell University

Paul Stodghill
USDA ARS Ithaca

David J. Schneider
USDA ARS Ithaca

Samuel Cartinhour
Department of Plant Pathology,
Cornell University

- Research funded by the US Department of Agriculture.

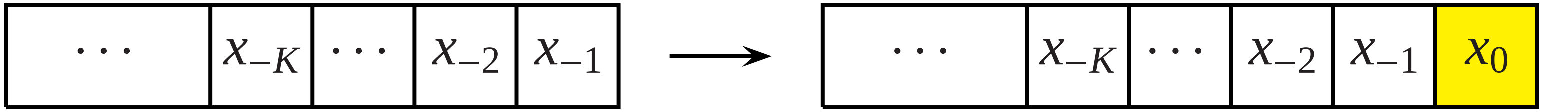
The Biological Sequence Segmentation Problem

- Two motivating problems:
  - HT segments (genomic islands) and lineage-specific segments (backbone) in bacterial DNA.
    - HT segments have different statistics from backbone.
    - Pathogenic genes frequently found near HT segment boundaries.
    - Gene-finding algorithms do not perform well in regions where statistics differ significantly from backbone.
    - Scoring problem even more severe for computational search of short regulatory elements.
  - Mesoscopic description of genome: 'Local' statistics vary along DNA sequence. Break long sequence into intermediate length segments, based on 'discernible' changes in statistics. Coarse-grained description.
- DNA polymerization along $5' \to 3'$ direction builds directionality into sequence. Biases in dinucleotide and codon frequencies. Model as Markov chains rather than Bernoulli chains with extended alphabets.

Markov chains

- State x_i of Markov chain at sequence position i can take on values in alphabet $\mathcal{S}$ of size S. Example. For DNA sequences, $\mathcal{S} = \{A, T, C, G\}$, and $S = 4$.

- Markov chains generated probabilistically. Existing subsequence extended



by attaching x_0 to end of subsequence with transition probability

$$p(x_0|x_{-1}x_{-2}\cdots x_{-K}).$$

- Markov chain of order K if $p(x_0|x_{-1}x_{-2}\cdots x_{-K'}) = p(x_0|x_{-1}x_{-2}\cdots x_{-K})$ for all $K' \geq K$.

- Transition probabilities can be organized into transition matrix

$$\mathbb{P} = [p_{\mathbf{t}s}], \quad s = 1, \ldots, S, \quad \mathbf{t} = t_1 \cdots t_K \in S^K.$$

- Equilibrium distribution $\boldsymbol{\pi} = (P_1, \ldots, P_k, \ldots, P_{S^K})$ such that $\boldsymbol{\pi}\mathbb{P} = \boldsymbol{\pi}$, P_k = probability of finding kth K-mer in stationary Markov chain.

Classification of Segmentation Schemes

- Matrix of segmentation schemes in literature:

	single−pass	recursive	local	global
sliding window average				
DNA walk				
dynamic programming				
hidden Markov model				

- All schemes rely on entropic measure of statistical dissimilarity, whether:
  - computed directly; or
  - in the form of inner product between quantized vectors of probabilities.

The Jensen-Shannon Divergence

- Given length-N sequence $\mathbf{x} = x_1 x_2 \cdots x_N$, $x_i = A, C, G, T$, assume composed of $M \geq 1$ Markov chains with boundaries at $i_1, \ldots, i_{M-1}$. M-segment sequence likelihood given by

$$P_M(\mathbf{x}; i_1, \ldots, i_{M-1}; \hat{\mathbb{P}}_1, \ldots, \hat{\mathbb{P}}_M) = \prod_{m=1}^{M} \prod_{\mathbf{t} \in S^K} \prod_{s=1}^{S} (\hat{p}^m_{\mathbf{t}s})^{f^m_{\mathbf{t}s}}; \qquad \hat{p}^m_{\mathbf{t}s} = \frac{f^m_{\mathbf{t}s}}{\sum_{s'} f^m_{\mathbf{t}s'}}.$$

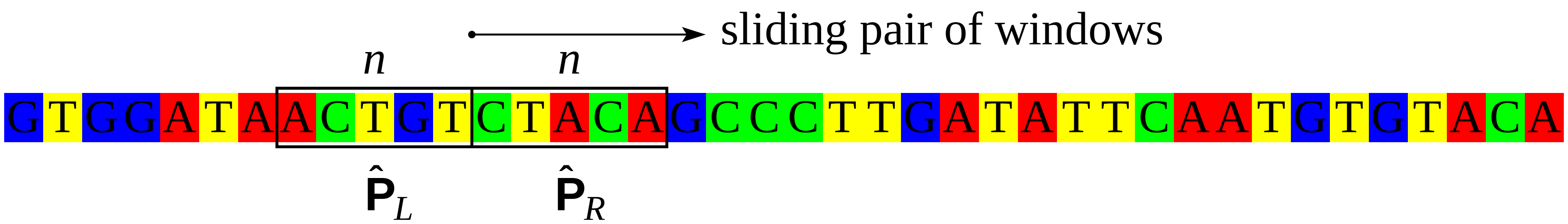
- Jensen-Shannon divergence

$$\Delta_M = \log \frac{P_M}{P_1} = -\sum_{\mathbf{t} \in S^K} \sum_{s=1}^{S} f_{\mathbf{t}s} \log \hat{p}_{\mathbf{t}s} + \sum_{m=1}^{M} \sum_{\mathbf{t} \in S^K} \sum_{s=1}^{S} f^m_{\mathbf{t}s} \log \hat{p}^m_{\mathbf{t}s};$$

$$f_{\mathbf{t}s} = \sum_{m=1}^{M} f^m_{\mathbf{t}s}, \qquad \hat{p}_{\mathbf{t}s} = \frac{f_{\mathbf{t}s}}{\sum_{s'=1}^{S} f_{\mathbf{t}s'}}$$

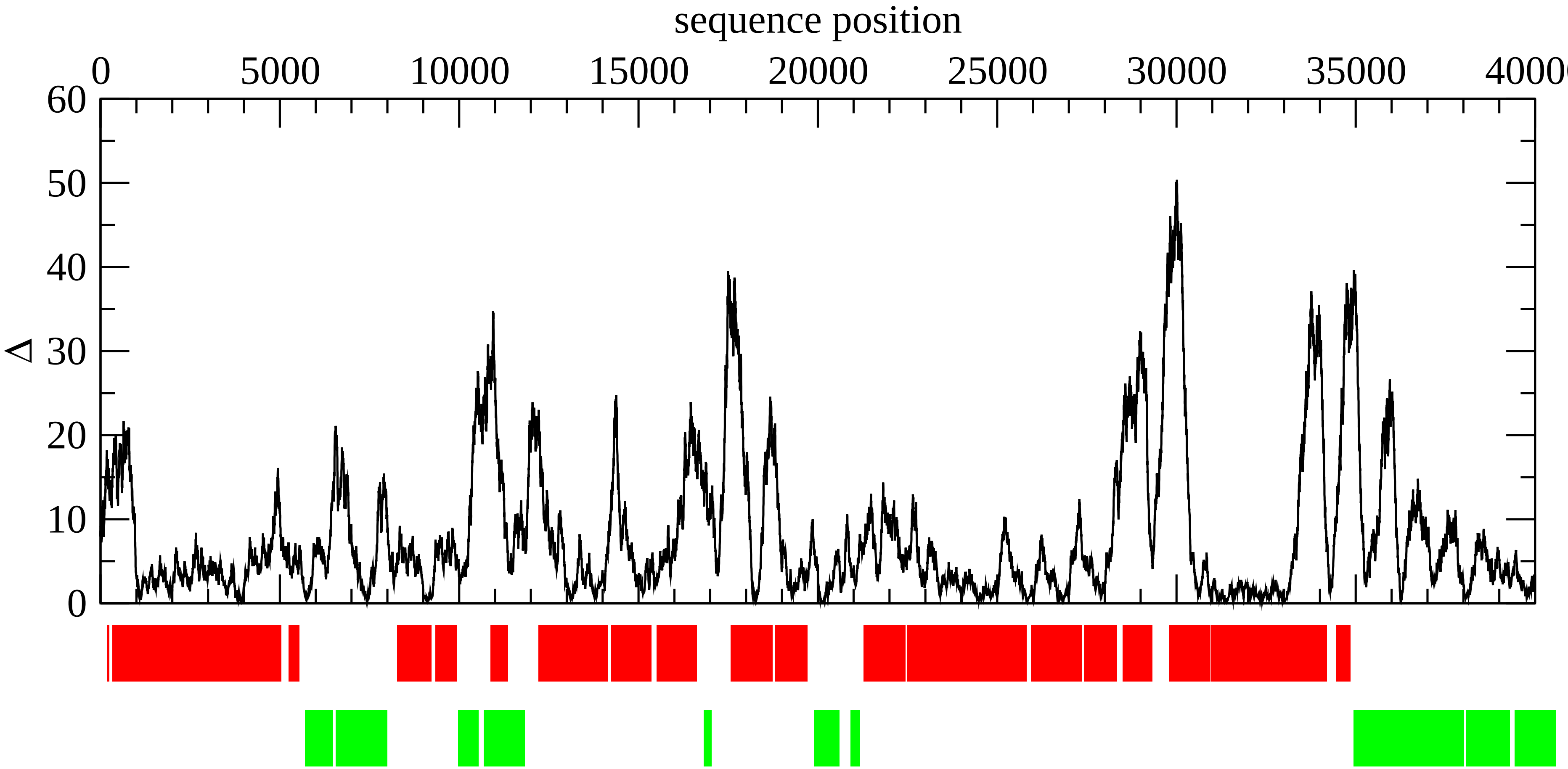
  is symmetric relative entropy providing quantitative measure of ‘goodness-of-fit’ of M-segment model over 1-segment model.

Segmentation with a Pair of Sliding Windows



- For a single sliding window of length n, spatial resolution decreases with n while statistical significance increases with n.
- Solution: To not compromise spatial resolution, use an adjoining pair of sliding windows, each of length n.
- Compute $\Delta_2(i)$ using $\hat{\mathbb{P}}_L$ in left window and $\hat{\mathbb{P}}_L$ in right window as function of sequence position i of centre of pair of windows.
- Segment boundaries appear as peaks in $\Delta_2(i)$. Strength of peak measure of statistical difference between the segments it separates.

Segmentation with a Pair of Sliding Windows

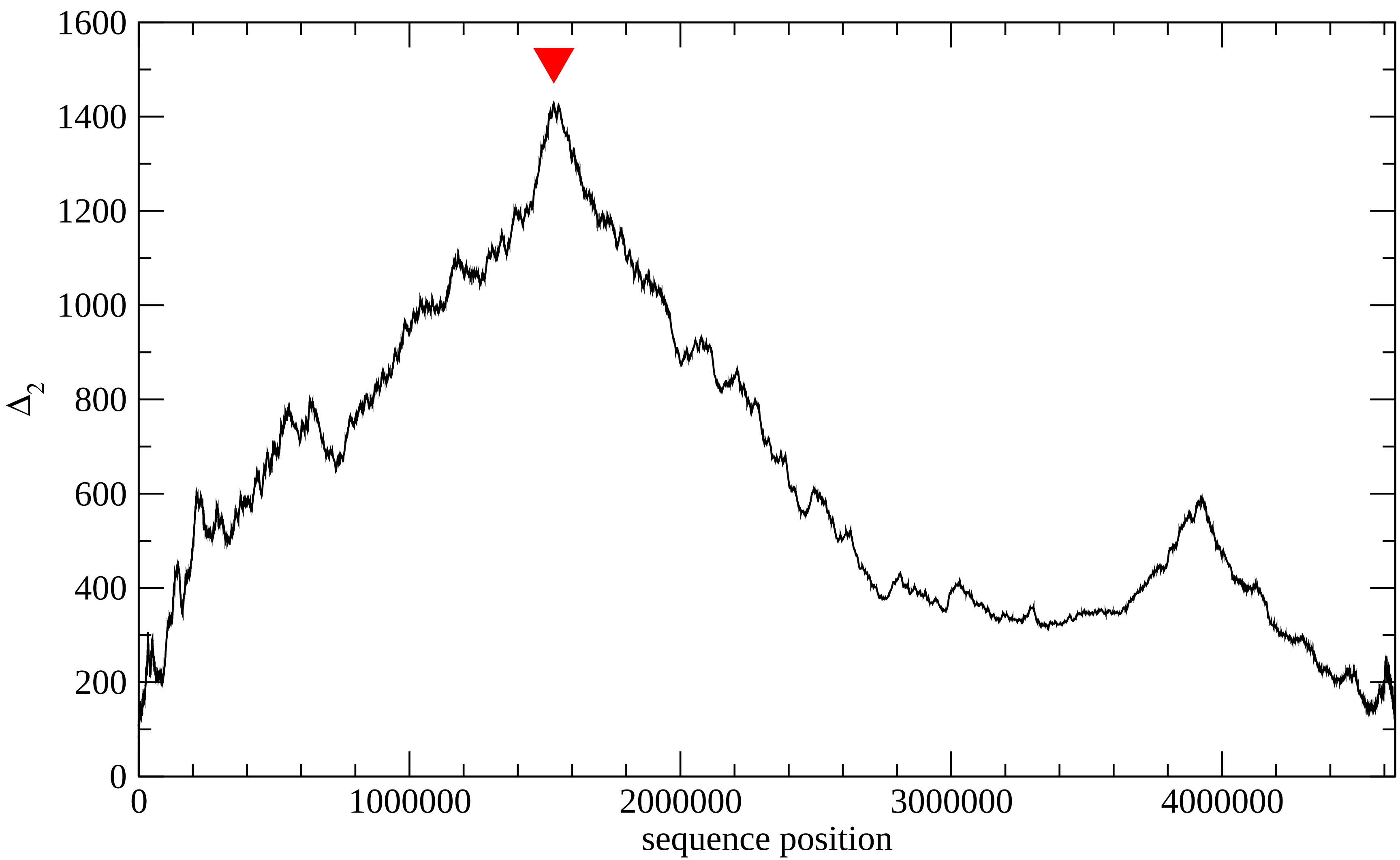


The interval $(0, 40000)$ in the *E. coli* K-12 MG1655 genome ($N = 4639675$), showing the $K = 0$ Jensen-Shannon divergence spectrum for $n = 1000$. Annotated genes on the positive (red) and negative (green) strands are shown below the graph.

Recursive Jensen-Shannon Segmentation

- STEP 1 (Segmentation):
  - Given sequence $\mathbf{x} = x_1 x_2 \cdots x_N$, compute 2-segment Jensen-Shannon divergence $\Delta_2(i)$ as function of cursor position i.
  - Find i^* such that $\Delta_2(i^*) = \max_i \Delta_2(i)$. The best 2-segment model for $\mathbf{x}$ is $\mathbf{x} = \mathbf{x}_L \mathbf{x}_R$, where $\mathbf{x}_L = x_1 \cdots x_{i^*}$ and $\mathbf{x}_R = x_{i^*+1} \cdots x_N$.
- STEP 2 (Recursion): Repeat STEP 1 for $\mathbf{x}_L$ and $\mathbf{x}_R$.
- STEP 3 (Termination): 1-segment model selected over 2-segment model if:
  - Hypothesis Testing: probability of obtaining divergence beyond observed Δ_2 greater than prescribed tolerance ϵ; or
  - Model Selection: information criterion (e.g. AIC, BIC) for 2-segment model greater than that for 1-segment model.

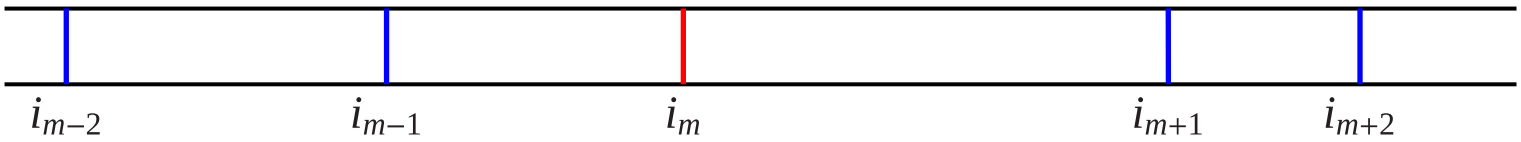
Recursive Jensen-Shannon Segmentation



Jensen-Shannon divergence spectrum of order $K = 3$ over the entire genome of *E. coli* K-12 MG1655 ($N = 4639675$ bp). The first segment boundary to be obtained in this first stage of recursive segmentation is shown by the red arrow.

Segmentation Optimization

- Two procedures to optimize segment boundary i_m if we are allowed to move only one segment boundary at a time:



  - First-order update: Compute $\Delta_2^m(i)$ for supersegment (i_{m-1}, i, i_{m+1}), and choose $i_m = i^*$, such that $\Delta_2(i^*) = \max_{i_{m-1}<i<i_{m+1}} \Delta_2(i)$, to be new position of segment boundary.
  - Second-order update: Compute $\Delta_2^{m-1}(i)$ for supersegment (i_{m-2}, i_{m-1}, i) and $\Delta_2^{m+1}(i)$ for supersegment (i, i_{m+1}, i_{m+2}), and choose $i_m = i^*$, such that

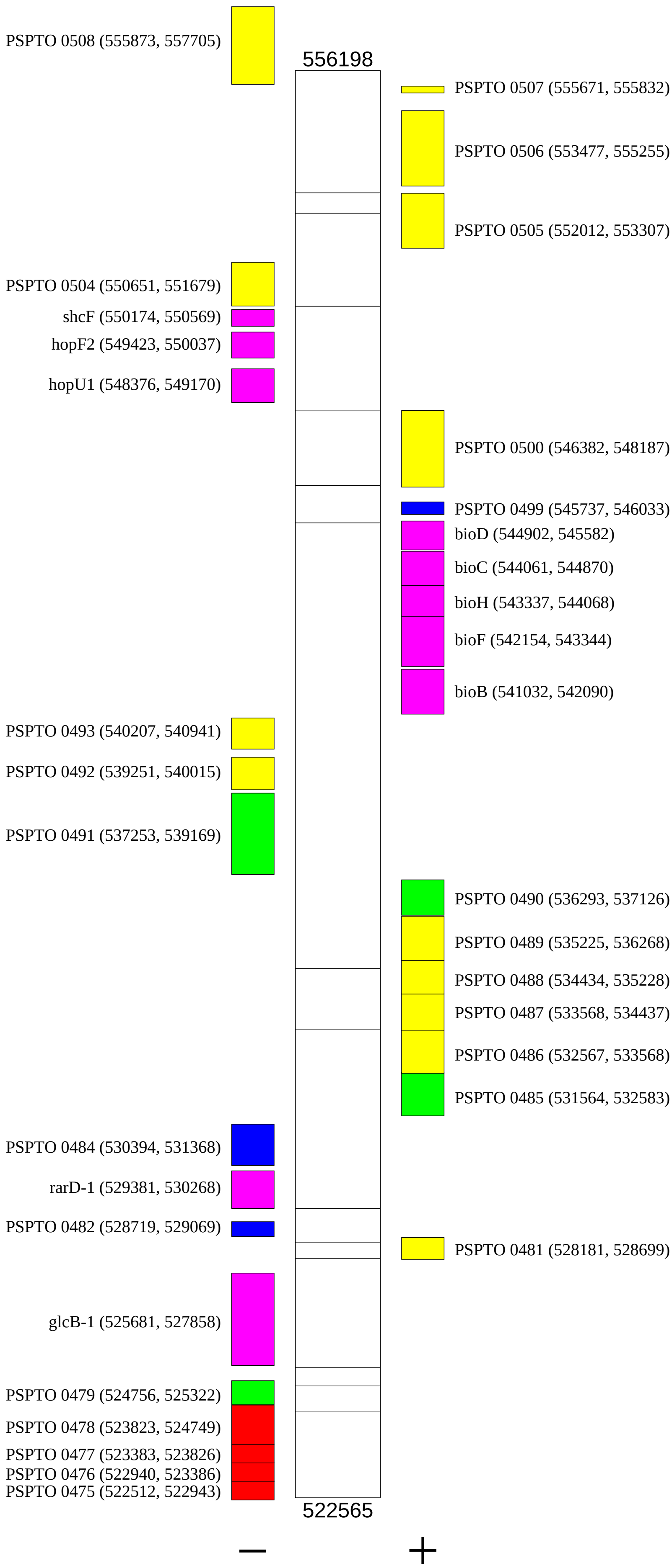
$$\Delta_2^{m-1}(i^*) + \Delta_2^{m+1}(i^*) = \max_{i_{m-1}<i<i_{m+1}} \left[\Delta_2^{m-1}(i) + \Delta_2^{m+1}(i)\right],$$

    to be new position of segment boundary.

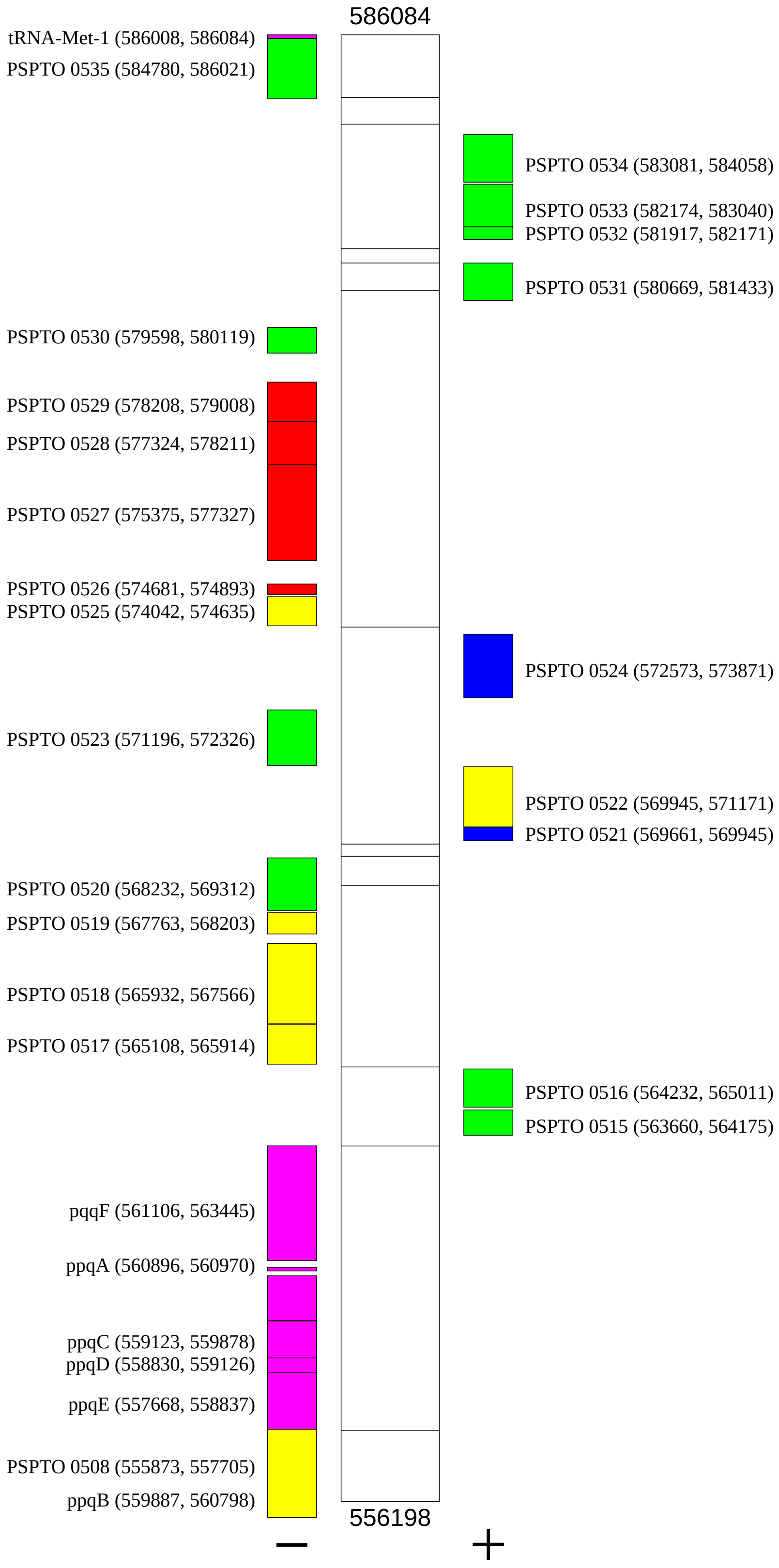
- Segment boundaries $\{i_m\}_{m=1}^M$ updated serially, or in parallel.

- Optimized recursive segmentation: Right after STEP 1 (Segmentation), optimize segmentation using first- or second-order update algorithm.

Optimized Recursive Jensen-Shannon Segmentation



Optimized Recursive Jensen-Shannon Segmentation



Conclusions & Further Works

- In conclusion, we have:
  - Developed segmentation scheme using a pair of sliding windows;
  - Developed optimization algorithms for recursive Jensen-Shannon segmentation scheme; and
- Further works:
  - Mean-field analysis of sliding window segmentation scheme: mean-field lineshape and match filtering;
  - Mean-field analysis of recursive segmentation scheme: identified problem of context sensitivity;
  - Developed new termination criterion based on intrinsic statistical fluctuations.
  - Incomplete segmentation misleading, cluster terminal segments instead to obtain coarser scale description of genome. E.g. to distinguish lineage-specific regions arising from HGT and the genetic backbone.
  - Multiple sequence clustering for comparative, phylogenetic studies.